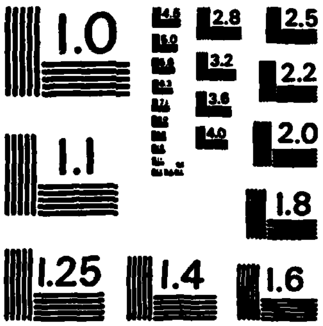
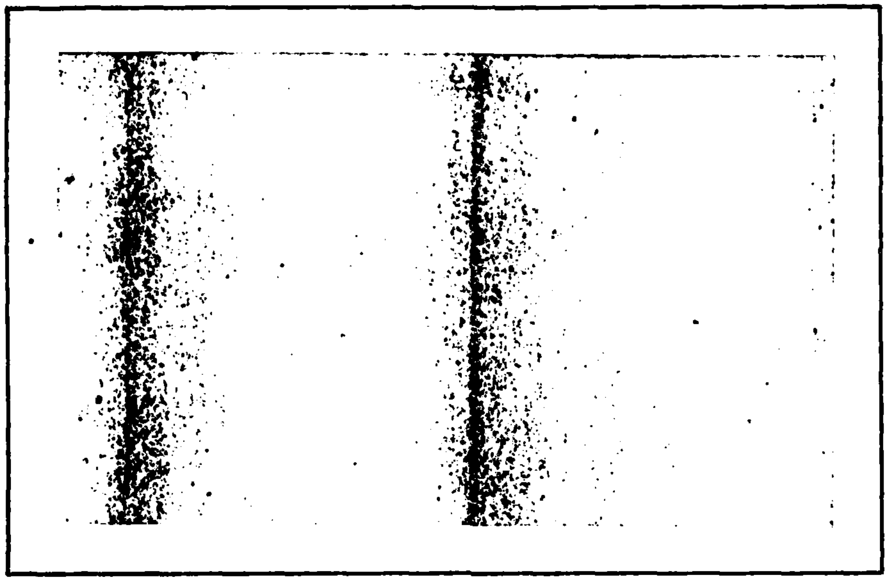
AD-A121 922 MULTIVARIATE DIRECTED GRAPHS IN STATISTICS(U) 1/1
CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS
S E FIENBERG OCT 82 TR-258 N00014-80-C-0637

UNCLASSIFIED F/G 12/1 NL

END
FILMED
DTIC


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

12



# DEPARTMENT
# OF
# STATISTICS

## Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

12

# MULTIVARIATE DIRECTED GRAPHS IN STATISTICS

by

Stephen E. Fienberg

Technical Report No. 258

Department of Statistics

Carnegie-Mellon University

Pittsburgh, PA 15213

October, 1982

The preparation of this paper was partially supported by the Office of Naval Research Contract N00014-80-C-0637 at Carnegie-Mellon University. Reproduction in whole or part is permitted for any purpose of the United States Government. This material will appear as an entry in S. Kotz and N.L. Johnson (eds.) *Encyclopedia of Statistical Sciences*. New York: Wiley.

## 1. UNIVARIATE DIRECTED GRAPHS

A direct graph consists of a set of g nodes and a set of directed arcs connecting pairs of nodes. Such graphs are natural mathematical representations of biological and social networks. They are also used in various other applications such as statistical geography and transportation networks, and in the study of disease contagion using acquaintance networks. For a social network the nodes of a graph may represent individuals, groups, or even organizations, and the arcs correspond to relationships or choices broadly interpreted to represent any type of binary relationship. It is customary (e.g. see Harary, Norman, and Cartwright, 1965) to use an incidence matrix representation of directed graphs. Thus, corresponding to each graph is an adjacency matrix, $x = (x_{ij})$, such that

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ chooses } j \\ 0 & \text{otherwise} . \end{cases} \qquad (1)$$

where $x_{ii} = 0$.

Holland and *Leinhardt (1975, 1979) and Frank (1971, 1981) summarize* the historical development of *random graphs*, for which the observed adjacency matrix is treated as the realization of a matrix random variable, X, which has a probability distribution on the set of all directed graphs with g nodes. Typically, the observed features of an empirically construc?ed directed graph are compared with the distribution of features that is generated by some random graph. This basic idea can be traced back in the social science literature to Moreno (1934).

One of the more interesting developments in the modelling of directed graphs is due to Holland and Leinhardt (1981), who begin by assuming independence of relationships amongst *pairs* of nodes or *dyads*. Their basic model can be represented in the form

$$
\begin{aligned}
\log \Pr[1-X_{ij})(1-X_{ji}) = 1] &= \lambda_{ij} , \\
\log \Pr[X_{ij}(1-X_{ji}) = 1] &= \lambda_{ij} + \alpha_i + \beta_j + \theta , \\
\log \Pr[(1-X_{ij})\, X_{ji} = 1] &= \lambda_{ij} + \alpha_j + \beta_i + \theta , \\
\log \Pr[X_{ij}\, X_{ji} = 1] &= \lambda_{ij} + \alpha_i + \alpha_j + \beta_i + \beta_j + 2\theta + \rho .
\end{aligned}
\qquad (2)
$$

The parameter $\lambda_{ij}$ is required for normalization purposes (each dyad must be in one of the four possible states), $\{\alpha_i\}$ and $\{\beta_j\}$ are effects that measure the productivity and attractiveness of the nodes, $\theta$ is a choice parameter, and $\rho$ is a measure "reciprocity." Note that model (2) is loglinear in structure. Holland and Leinhardt present iterative methods for maximum likelihood estimation for the parameters in this model (see the discussion of the estimation of parameters in loglinear models for categorical data in CONTINGENCY TABLES), and Fienberg and Wasserman (1981a) provide an alternative approach based on a simple transformation of the data and the use of the method of *iterative proportional fitting**. They also suggest several generalizations of the Holland-Leinhardt model where, for example, the parameter $\rho$ in expression (2) is replaced by

$$\rho_{ij} = \rho + \rho_i + \rho_j , \qquad (3)$$

where $\Sigma\rho_i = 0$, and demonstrate how the parameters of this model can also be estimated by iterative proportional fitting.

Two outstanding theoretical statistical problems in connection with the Holland and Leinhardt univariate model and its generalizations are (i) the lack of an appropriate asymptotic framework for inference (see the discussions in Fienberg and Wasserman (1981b) and Haberman (1981)) which is needed to carry out goodness-of-fit tests, and (ii) the need for alternative models which allow for dyadic dependence and include the Holland-Leinhardt model as a special case.

## 2. MULTIVARIATE DIRECTED GRAPHS

A *multivariate directed graph* is simply a collection of univariate directed graphs with the same g nodes. (The term *multi-graph* is also in wide-spread use.) If there are R such univariate graphs, then we represent the multivariate graph by the collection of adjacency matrices for the R univariate graphs, $\{x_1, x_2,..., x_R\}$. We may think of the R graphs as representing either R different types relationships amongst the g nodes, or the same relationship at R different points in time. In either case, we wish to think of an observed multivariate graph as a realization of a random multivariate graph $X = \{X_1, X_2,..., X_R\}$.

In the univariate situation we saw that each dyad had four possible realizations:

$$
\begin{aligned}
(1,1) &: \text{ arcs in both directions },\\
(1,0) \text{ or } (0,1) &: \text{ arc in one direction },\\
(0,0) &: \text{ no arc }.
\end{aligned}
$$

Now each dyad has $2^{2R}$ possible realizations.

Fienberg, Meyer, and Wasserman (1981) have proposed a class of loglinear models for random multivariate directed graphs, that generalize some aspects of the Holland-Leinhardt model to the multivariate case. By sacrificing the node-level parameters, $\{\alpha_i\}$ and $\{\beta_j\}$, associated with each univariate graph, these models incorporate not only *reciprocity* effects for dyadic patterns of the form:

$$X \underset{\text{Relation } r}{\longleftrightarrow} Y \, ,$$

but also *exchange* effects for patterns of the form:

$$X \underset{\text{Relation } r_2}{\overset{\text{Relation } r_1}{\rightleftarrows}} Y \, ,$$

and *multiplex* choice effects for patterns of the form:

$$\begin{tikzcd} X \arrow[r, shift left=2, "\text{Relation } r_1"] \arrow[r, shift right=2, "\text{Relation } r_2"'] & Y, \end{tikzcd}$$

as well as multivariate generalizations of these effects.

Although there are $2^{2R}$ possible dyadic realizations we only get to observe

$$2^R + 2^R(2^R - 1)/2$$

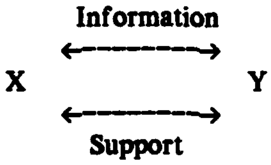
states when the nodes lose their individual identities. We can still summarize the data in the adjacency matrices of the multivariate graph by counting every dyad twice, once from the perspective of each node. As a consequence we end up with a $2^{2R}$ table of counts, with the $2^R$ cells corresponding to reciprocal arcs on each relation (both present or both absent) containing double the actual number of dyads, and each of the remaining $2^{R-1}(2^R-1)$ patterns yielding two symmetrically placed duplicate counts in the table.

Fienberg, Meyer, and Wasserman (1981) show how fitting a simple affine translation of a loglinear model for the $2^R + 2^{R-1}(2^R-1)$ counts corresponds to fitting standard loglinear models to the $2^{2R}$ table of duplicated and doubled counts.

## 3. TWO EXAMPLES

Holland and Leinhardt (1981) illustrate their univariate model on data collected by Sampson (1969) who spent a year observing monks in an American Monastery. Sampson measured both negative and positive relationships on four dimensions at five different points in time. The same 18 monks were interviewed at three of these time points. Thus the data can be represented in the form of an $R = 4 \times 2 \times 3 = 24$ variate directed graph involving 18 nodes. Holland and Leinhardt analyze only a single relationship from this data-set.

Galaskiewicz and Marsden (1978) and Fienberg, Meyer, and Wasserman (1981) describe data from a study of the formal organizations in a small midwest U.S. community of 32,000 persons referred to by the pseudonym "Towertown." They focus their analyses on a subset of 73 organizations and their links on three relations: (1) information, (2) money, and (3) support. Thus the original data take the form of three 73×73 adjacency matrices, but the analyses focus on a summary of these in the form of a $2^6$ table of counts of pairs of organizations. The full adjacency matrices are availabe in Fienberg and Galaskiewicz (1982). The most substantial estimated effects in the loglinear models fitted by Fienberg, Meyer, and Wasserman are associated with choices ($\theta$'s), reciprocity ($\rho$'s), and a multiplex-reciprocity effect associated with the dyadic pattern:



## 4. SOME RELATED STATISTICAL APPROACHES

In a pair of related papers, White, Boorman and Breiger (1976) and Boorman and White (1976) proposed a method, labelled as *blockmodelling*, for the analysis of data in the form of multivariate directed graphs. A *blockmodel* for a network consists of a partition of the nodes into blocks of structural equivalent nodes (i.e. ones which relate in the same way to all other nodes in the network), and corresponds to a deterministic rather than a stochastic model. Unfortunately, few directed graphs yield exactly to such blockmodels, and substantive social science theory does not always suggest appropriate partitions. Thus White, Boorman, and Brieger suggested the use of a statistical-like approach to the search for an "acceptable" block model of a particular form, and they demonstrate their approach on Sampson's monastery data described above.

Breiger, Boorman and Arabie (1976) describe a more general search procedure for a block model structure, based on *hierarchical clustering** methods, and apply their method to a study of directorship interlocks in American industry. These methods are closely related to other exploratory statistical procedures for row-column permutations of a matrix, such as nonmetric *multidimensional scaling** (see Arabie, Boorman, and Levitt (1978)).

Major drawbacks of blockmodel methods include: (i) their inexplicit use of formal parametric models, (ii) the use of arbitrary criterion functions for the choice of partitions, (iii) the inability to distinguish actual structure from chance variation. Their major advantage is that they provide an explicit model for the pattern of responses, which many sociometricians find very useful for thinking about sociological theory (see Light and Mullins (1979)).

## REFERENCES

Arabie, P., Boorman, S.A., and Levitt, P.R. (1978). *J. Math. Psych.*, 17, 21-63.

Boorman, S.A. and White, H.C. (1976). *Amer. J. Sociol.*, 81, 1384-1446.

Breiger, R.L., Boorman, S.A., and Arabie, P. (1975). *J. Math. Psych.*, 12, 328-383.

Fienberg, S.E. and Galaskiewicz, J. (1982). In *Data* (D. Andrews and A. Herzberg, eds.), Springer Verlag, New York, in press.

Fienberg, S.E., Meyer, M.M. and Wasserman, S.S. (1981). In *Interpreting Multivariate Data* (V. Barnett, ed.) New York: Wiley, 289-306.

Fienberg, S.E. and Wasserman, S.S. (1981a). *Sociological Methodology 1981*, 156-192.

Fienberg, S.E. and Wasserman, S.S. (1981b). *J. Amer. Statist. Assoc.*, 76, 54-57.

Frank, O. (1971). *Statistical Inference in Graphs.* Swedish Research Institute of National Defense, Stockholm.

Frank, O. (1981). *Sociological Methodology 1981*, 110-155.

Galaskiewicz, J. and Marsden, P.V. (1978). *Soc. Sci. Res.*, 7, 89-107.

Haberman, S.J. (1981). *J. Amer. Statist. Assoc.*, 76, 60-62.

Harary, F., Norman, R.Z., and Cartwright, D. (1965). *Structural Models: An Introduction to the Theory of Directed Graphs.* Wiley, New York.

Holland, P.W. and Leinhardt, S. (1975). *Sociological Methodology 1976.*

Holland, P.W. and Leinhardt, S. (1979). In *Perspectives on Social Network Research* (P.W. Holland and S. Leinhardt, eds.), 63-83.

Holland, P.W. and Leinhardt, S. (1981). *J. Amer. Statist. Assoc.*, 76, 33-50.

Light, J.M. and Mullins, N.C. (1979). In *Perspectives on Social Network Research* (P.W. Holland and S. Leinhardt, eds.), 85-118.

Moreno, J.L. (1934). *Who Shall Survive?* Nervous and Mental Disease Publishing Co., Washington.

Sampson, S.F. *Crisis in a Cloister.* Ph.D. Thesis, Department of Sociology, Cornell University.

White, H.C., Boorman, S.A. and Breiger, R.L. (1976). *Amer. J. Sociol.*, 81, 730-780.

## ACKNOWLEDGEMENT

The preparation of this entry was supported in part by Contract N00014-80-C-0637 from the Office of Naval Research to Carnegie-Mellon University.

## RELATED ENTRIES

CONTINGENCY TABLES

HIERARCHICAL CLUSTERING

ITERATIVE PROPORTIONAL FITTING

MULTIDIMENSIONAL SCALING

| | | |
|---|---|---|
| **REPORT DOCUMENTATION PAGE** | | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM |
| 1. REPORT NUMBER<br>Technical Report No. 258 | 2. GOVT ACCESSION NO.<br>A121922 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>Multivariate Directed Graphs in Statistics | | 5. TYPE OF REPORT & PERIOD COVERED<br>6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Stephen E. Fienberg | | 8. CONTRACT OR GRANT NUMBER(s)<br>N00014-80-C-0637 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of Statistics<br>Carnegie-Mellon University<br>Pittsburgh, PA 15213 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Contracts Office<br>Carnegie-Mellon University<br>Pittsburgh, PA 15213 | | 12. REPORT DATE<br>October, 1982<br>13. NUMBER OF PAGES<br>8 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Contingency tables, Hierarchical clustering, Iterative proportional fitting, Loglinear models, Multidimensional scaling.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Directed graphs are natural mathematical representations of biological and social networks, and are used in other areas of application such as statistical geography, transportation, and epidemiology. This article reviews some of the recent statistical literature on the analysis of directed graphs, especially in the multivariate case.

END

FILMED

1-83

DTIC